Specification and assessment of methods supporting the development of neural networks in medicine

Summary

1 Scope of the dissertation

This doctoral dissertation presents methods and techniques that may expedite application of neural networks in medicine. Research on neural networks started as a branch of neurology. A neural network consists of a set of interconnected nodes (neurons). Each node works like a junction between "nerve" paths. The neuron receives a number of inputs and produces an activation. The activation of a neuron is functionally dependent on the input signals the neuron receives. Each input signal is modified by a weight. Since their introduction by McCulloch and Pitts in 1943, neural networks migrated to cognitive science, artificial intelligence, statistical regression and decision theory, signal processing and other engineering disciplines. Neural networks have been developed for a large number of applications in economy, computer science, telecommunication and medicine. Chapter 1 contains a brief overview of neural networks that have been developed for clinical decision support.

Clinical application of neural networks is problematic because of their black-box nature. It is very difficult to assess the knowledge encoded in the weights of a trained neural network as it constitutes a nonlinear mapping between the feature (input) space and the class (output) space. In the dissertation, different techniques that characterize the properties of a trained neural network are suggested. Thereby, development and verification/validation of neural networks is expedited. To enhance the application of neural networks, the topic of missing data is also addressed.

2 A neural network performs a mapping

The most general notion of a classifier is a mathematical mapping from an n-dimensional input space to a c-dimensional output space, $N: \mathbb{R}^n \rightarrow \mathbb{R}^c$, where n is the number of features or attributes[1] and c the number of classes to be discriminated. Neural networks can process combinations of qualitative and quantitative data. The c classes can be decisions such as diagnoses or therapies. The mapping is performed by a neural network, more specifically by the weighted connections between the input, hidden and output layers. During training of a neural network, the weights are adapted to minimize a function that measures the difference between the correct output of the learning cases and the output from the neural network.

[1]The terms feature and attribute are used interchangeably.

3 Assessing the output of a classifier

In chapter 2, metrics are defined that characterize the performance of a trained neural network. The performance is measured by letting the neural network classify a set of test cases of which the true class label is known. A contingency table (confusion matrix) is used to characterize the performance of a neural-net classifier. Existing metrics that characterize different properties of the neural-net classifier are discussed. Also some new metrics are introduced. Although these metrics are defined for a neural-net classifier, they can be applied to other classifiers of which the results can be characterized by a contingency table. The metrics include *correctness* – the fraction of correctly classified cases – and *coverage* – the fraction of cases to which the neural network can assign a class label. The misclassified cases are characterized by the metrics for *bias* and *dispersion*. Standard errors and confidence intervals for some of the metrics are specified. The usefulness of the metrics is explored in a set of experiments in which neural networks are trained for classification of thyroid disorders.

4 Assessing the importance of attributes for a classifier

The chapters 3 and 4 address how to assess the contribution of individual attributes to the performance of a neural-net classifier. The motivation for performing attribute assessment is twofold. First, one wants to obtain insight into which attributes are important for assigning a correct class label to one or more cases. Secondly, one needs a criterion to rank the attributes according to their contribution to the performance of the neural-net classifier, before unimportant attributes can be pruned.

In chapter 3, different approaches to attribute selection such as forward, backward and Branch-and-Bound search are discussed. It is argued that backward search is a suitable selection strategy. Based on a mathematical analysis of a minimal error-rate classifier, a metric for the *discriminative power* of an attribute is introduced. This metric is used as a criterion to rank the attributes for each case in the test set. The ranks for each attribute are summed over all cases. The summed ranks are compared using Friedman's two-way analysis of variance. Attributes with a high average rank are unimportant for the neural network whereas attributes with a low average rank have the most influence on the classification performance. The usefulness of this approach is assessed in a number of experiments with artificial classification problems. The experiments indicated that the approach ranks the attributes correctly, when applied on classifiers trained with independent attributes as well as on classifiers trained with dependent attributes. The approach is also used in an application to identify attributes that are important for discriminating four different types of texture in radiographs of focal bone lesions.

In chapter 4, a mathematical framework is developed in which four different feature measures are derived from a minimal error-rate classifier. Each measure allows one to compute a lower bound for the *marginal contribution* of a feature to the performance of a statistical classifier. These measures characterize the *influence* and

replaceability of a feature. Influence is the probability that a feature can possibly change the class label of a case while the other feature values are kept fixed. Replaceability is the expected decrease in performance when a feature value is substituted by the conditional mean of the feature.

Each feature measure is made operational by a feature metric. Computation of three of the four metrics requires the identification of the attribute-conditional decision boundaries. The decision boundaries for a given feature depend on the values of the other $n-1$ features and have to be identified in each case. The boundaries are identified with a piecewise polynomial approximation which is based on a Taylor expansion of the output of a neural-net classifier as a function of the given feature.

A pruning method called *LMS-pruning* is introduced. A feature is LMS-pruned by removing the links that connect the input node of the feature with the hidden nodes and changing the weights that connect the remaining features with the hidden nodes. The weights are modified such that the pruned neural network classifies the training cases identically to a network based on n feature values with the value of the pruned feature replaced by its expected value.

In experiments with artificial classification tasks, the four metrics are compared with respect to their ability to rank the features. These experiments indicate that replaceability is the best ranking criterion. The experiments showed that for neural-net classifiers with a performance close to the minimal error rate, LMS-pruning a feature resulted in a pruned network with a performance that remains close to the maximal (Bayesian) correctness.

5 Estimation of missing data

In chapter 5, a method for iterative estimation of missing data is suggested. Statistical classifiers such as neural networks require all inputs to be able to assign a class label to a case. This impedes application of such classifiers in environments where incomplete data frequently occur. Different approaches to estimate missing data such as the EM-algorithm and Multiple Imputation are discussed. To cope with some drawbacks of these two methods, it is suggested to use an auto associator neural network in recurrent mode to estimate missing values. The properties of an auto associator that is trained with complete cases is analyzed. Subsequently, it is suggested to use the auto associator in recurrent mode to estimate missing values. The conditions that ensure convergence of the recurrent auto associator are derived. It is proven that convergence is only possible when the number of hidden nodes of the auto associator is smaller than or equal to the number of observed values in an incomplete case.

The recurrent auto associator is embedded in the Recurrent Expectation Maximization (REM) algorithm, an iterative approach for estimating missing values in a set of cases. In a set of experiments, the residual variance of predictions made by the recurrent auto associator is compared with the residual variance obtained using multivariate linear regression. Also the REM and EM-algorithms are compared with respect to their ability to estimate missing values (residual variance) and to estimate

the covariance matrix from the incomplete sample. The experiments indicate that the recurrent auto associator results in poorer estimates of the missing values than multivariate regression. The REM-algorithm estimates the covariance matrix slightly worse than the EM-algorithm when the data are fairly correlated and all variables have identical variances. However, the REM-algorithm gives an indication of those combinations of variables with missing and observed values in which the missing data will be predicted poorly. Leaving out such cases leads to an improvement in the estimation of the covariance matrices by the REM-algorithm.

6 Classification from noisy attributes

In chapter 6, the influence of measurement noise on the classification of a case is analyzed. Based on ideas of Brender *et al.*, a quality measure called robustness is specified. The robustness of a classification is the probability that the class label assigned to the case would not be different from the classification based on the (unknown) true attribute values. It is assumed that the measurement noise is Gaussian with a zero mean and uncorrelated with the attributes. A formula for the robustness of a classification is specified.

In practice, it is difficult to estimate the robustness of a classification when the probability density function of the uncontaminated attributes is unknown. Therefore, two approximations are suggested. The bias introduced by these two approximations is analyzed for the special situations where an attribute comes either from a unimodal or a bimodal distribution and is to be classified into one of two classes.

A simulation experiment illustrates how often an attribute has to be remeasured to achieve a robust classification (the measurements are averaged, which reduces the influence of the measurement noise on the attribute value). It is clear that remeasuring a (noisy) attribute makes sense when only a few remeasurements are required to ensure a classification with a sufficiently high robustness. When, however, the robustness of a classification becomes too low, the number of measurements that are necessary to obtain a more accurate estimate of the attribute values becomes very high. The notion of *remeasuring intervals* is introduced. Such intervals indicate when remeasuring an attribute makes sense.

7 General conclusion

The methods and techniques developed in this dissertation are explored in a set of experiments. In chapter 7, it is discussed to which extent these methods and techniques may support development, verification and validation of neural networks. The possibility of applying knowledge-based systems in general and neural networks in particular in the clinic is discussed as well. It is argued that introduction of such systems in clinical practice interferes directly with the work processes of physicians. One can expect that such systems will have their largest potential in low-level information processing. It is an issue for further research to investigate the value of the presented methods and techniques in the development and evaluation of neural networks for clinical application.

Samenvatting

1 Onderwerp van het proefschrift

Dit proefschrift presenteert een aantal methoden en technieken die de bruikbaarheid van neurale netwerken in de kliniek kunnen vergroten. Onderzoek op het gebied van (kunstmatige) neurale netwerken begon binnen de neurologie. Een neuraal netwerk bestaat uit een aantal met elkaar verbonden neuronen (knopen). Elke kunstmatige neuron werkt als een knooppunt van "zenuwbanen". Het neuron ontvangt een aantal ingangssignalen, hetgeen resulteert in een bepaalde activatie van het neuron. Deze activatie is functioneel afhankelijk van de invoer die het neuron ontvangt. Elk ingangssignaal wordt gemodificeerd door een gewicht. Sinds hun introductie door McCulloch en Pitts in 1943 vindt onderzoek naar kunstmatige neurale netwerken plaats binnen de cognitieve wetenschap, kunstmatige intelligentie, statistische regressie en beslissingstheorie, signaalverwerking en andere technische wetenschappen. Neurale netwerken zijn ontwikkeld voor een groot aantal toepassingen in de economie, informatica, telecommunicatie en geneeskunde. Hoofdstuk 1 bevat een kort overzicht van neurale netwerken die zijn ontwikkeld voor klinische beslissingsondersteuning.

De klinische toepassing van neurale netwerken is problematisch vanwege hun 'black-box' karakter. Het is zeer moeilijk vast te stellen welke kennis gecodeerd is in de gewichten van een getraind neuraal netwerk, omdat het netwerk een nietlineaire afbeelding vormt tussen de invoer (kenmerk) ruimte en de uitvoer (klassen) ruimte. In dit proefschrift wordt een aantal technieken gepresenteerd die de eigenschappen van een getraind neuraal netwerk kunnen karakteriseren. Hiermee wordt de ontwikkeling, verificatie en validatie van neurale netwerken ondersteund. Om de toepasbaarheid van neurale netwerken in de medische praktijk te vergroten wordt tevens een methode geïntroduceerd voor het schatten van ontbrekende data.

2 Een neuraal netwerk verricht een afbeelding

De meest algemene beschrijving van een classificator is een wiskundige afbeelding van een n-dimensionale kenmerkruimte naar een c-dimensionale klassenruimte, N: $\mathbb{R}^n \rightarrow \mathbb{R}^c$. Neurale netwerken kunnen combinaties van kwalitatieve en kwantitatieve kenmerken verwerken. De c klassen kunnen bijvoorbeeld diagnoses of therapieën zijn. De afbeelding wordt verricht door het neuraal netwerk, meer specifiek door de gewogen connecties tussen de invoer-, de verborgen- (hidden) en de uitvoerlaag. De gewichten worden in het leerproces zodanig aangepast dat het verschil tussen de geproduceerde en de gewenste uitvoer voor een aantal leercasus minimaal is.

3 Beoordeling van de prestaties van een classificator

In hoofdstuk 2 worden metrieken gedefinieerd die de prestaties van een getraind neuraal netwerk karakteriseren. De prestaties worden gemeten door testcasus, waarvan het juiste klassenlabel bekend is, aan te bieden aan een neuraal netwerk. De prestaties van een netwerk worden in een kruistabel weergegeven. Bestaande metrieken die de waarden in een kruistabel omzetten in kengetallen worden bediscussieerd en nieuwe metrieken worden geïntroduceerd. Alle metrieken zijn gedefinieerd voor een neuraal netwerk, maar ze zijn ook geschikt om de prestaties van andere soorten classificatoren te meten, indien de resultaten hiervan kunnen worden weergegeven in een kruistabel. Besproken worden onder andere de metrieken voor *correctheid* – de fractie correct geclassificeerde casus – en *dekking* (coverage) – de fractie van casus waaraan de classificator een klassenlabel kan toekennen. De incorrect geclassificeerde casus worden gekarakteriseerd met de metrieken *bias* en *dispersie*. Formules voor standaard fouten en betrouwbaarheidsintervallen worden voor een aantal van de metrieken gegeven. De toepasbaarheid van de metrieken is in een aantal experimenten onderzocht. In deze experimenten zijn neurale netwerken getraind om patiënten met schildklierafwijkingen te classificeren op basis van de concentraties van een aantal hormonen in het bloed.

4 Beoordeling van de kenmerken van een classificator

In de hoofdstukken 3 en 4 worden twee methoden beschreven die de bijdrage meten van individuele kenmerken aan de prestaties van een neuraal netwerk. Het doel van kenmerkanalyse is tweeledig. Ten eerste wil men graag weten welke kenmerken belangrijk zijn voor het toekennen van het correcte klassenlabel aan casus. Ten tweede is een criterium nodig om de kenmerken te kunnen rangschikken op basis van hun bijdrage aan de prestatie van het neuraal netwerk voordat onbelangrijke kenmerken die weinig bijdragen verwijderd kunnen worden.

In hoofdstuk 3 worden verschillende zoekstrategieën bediscussieerd, zoals forward en backward search en het Branch-and-Bound algoritme. Hieruit volgt dat backward search een geschikte zoekstrategie is. Op basis van een wiskundige analyse van een minimale-fout classificator wordt een metriek voor het discriminerend vermogen van een kenmerk gedefinieerd. Deze metriek dient als criterium om de n kenmerken te sorteren voor elke casus afzonderlijk. De rangordes worden per kenmerk opgeteld over alle casus. De n sommen worden vervolgens met elkaar vergeleken met behulp van Friedman's tweezijdige variantieanalyse. Kenmerken met een hoge gemiddelde rang zijn onbelangrijk; kenmerken met een lage gemiddelde rang hebben een hoge bijdrage aan het discriminerend vermogen van de classificator. De bruikbaarheid van de voorgestelde methode wordt getoetst in een aantal experimenten met kunstmatige classificatieproblemen. De experimenten laten zien dat de methode de n kenmerken op de juiste manier ordent, ongeacht of deze binnen de klassen afhankelijk zijn of niet. De bruikbaarheid van deze methode wordt gedemonstreerd aan de hand van een toepassing waarbij vier soorten textuur moeten worden onderscheiden in röntgenbeelden van botten die mogelijk een tumor bevatten.

In hoofdstuk 4 wordt een mathematisch kader geschetst waarbinnen vier verschillende maten voor de bijdrage van een kenmerk zijn afgeleid voor een minimale-fout classificator. Elke maat maakt het mogelijk een ondergrens te berekenen voor de *marginale bijdrage* van een kenmerk aan de prestaties van een statistische classificator. Deze maten karakteriseren samen de *invloed* en *vervangbaarheid* van een kenmerk. Invloed geeft de waarschijnlijkheid aan dat een waarde van een kenmerk kan bepalen welke klassenlabel wordt toegekend aan een casus casus gegeven de waarden van de overige kenmerken die constant worden gehouden. De vervangbaarheid geeft de te verwachte daling in correctheid aan, wanneer de kenmerkwaarden worden vervangen door hun conditionele gemiddelde.

Elke kenmerkmaat wordt geoperationaliseerd door een kenmerkmetriek. Om deze metrieken te kunnen berekenen moet men de kenmerk-conditionele klassengrenzen bepalen. De klassengrenzen voor een bepaald kenmerk zijn afhankelijk van de waarden van de $n-1$ andere kenmerken en daarom moeten de grenzen voor elke casus afzonderlijk bepaald worden. De grenzen worden bepaald met een polynomische benadering van het uitvoer van het neuraal netwerk als functie van het gekozen kenmerk.

Een methode om invoerneuronen weg te snoeien – *LMS-pruning* – wordt geïntroduceerd. LMS-pruning van een kenmerk bestaat uit het verwijderen van de connecties tussen het invoerneuron overeenstemmend met het te verwijderen kenmerk en de hidden neuronen, en het modificeren van de gewichten tussen de resterende kenmerken en de hidden neuronen. Deze gewichten worden zodanig aangepast dat het gesnoeide netwerk de leercasus classificeert als een netwerk gebaseerd op alle n kenmerkwaarden waarbij de gesnoeide kenmerkwaarde is vervangen door zijn conditionele gemiddelde.

De ordening van de kenmerken op basis van de vier kenmerkmaten wordt vergeleken met de correcte rangschikking van elk kenmerk in experimenten met kunstmatige classificatietaken. Deze experimenten tonen aan dat de vervangbaarheid het beste ordeningscriterium is. De experimenten laten tevens zien dat voor neurale netwerken die bijna zo goed presteren als de minimale-fout classificator, LMS-pruning van een kenmerk een gesnoeid netwerk oplevert met een prestatie die dicht bij de maximale (Bayesiaanse) correctheid komt.

5 Het schatten van ontbrekende data

In hoofdstuk 5 wordt een methode voorgesteld voor het iteratief schatten van ontbrekende data. Voor statistische classificatoren, zoals neurale netwerken, zijn alle invoergegevens nodig om een klassenlabel te kunnen toekennen aan een casus. Dit maakt het moeilijk deze classificatoren toe te passen in situaties waarin vaak gegevens ontbreken. Verschillende methoden voor het schatten van ontbrekende gegevens, zoals het EM-algoritme en Multiple Imputatie worden bediscussieerd. Om een aantal nadelen van deze methoden te vermijden wordt voorgesteld een zelf-associerend neuraal netwerk in recurrent mode te gebruiken en daarmee de ontbrekende gegevens te schatten. Eerst wordt de situatie geanalyseerd waarin een zelf-associerend netwerk getraind is met een set van complete casus. De condities

waaronder het recurrent zelf-associerende netwerk convergeert worden afgeleid. Uit het bewijs blijkt dat convergentie alleen plaats kan vinden wanneer het aantal hidden neuronen van de auto associator kleiner of gelijk is aan het aantal geobserveerde waarden in een incomplete casus.

De auto associator in recurrent mode maakt deel uit van de Recurrent Expectation Maximization (REM) algoritme, hetgeen een iteratieve methode is voor het schatten van ontbrekende data in een database. In een aantal experimenten wordt de residuele variantie behaald met het zelf-associerend neuraal netwerk in recurrent mode vergeleken met de residuele variantie die behaald wordt met multiple regressie. De REM en EM-algoritmen worden vergeleken voor wat betreft het schatten van ontbrekende data en het schatten van de covariantiematrix van de incomplete database. Volgens de experimenten levert de auto associator in recurrent mode slechtere schattingen op van de ontbrekende data dan multiple regressie. Het REM-algoritme resulteert in iets slechtere schattingen van de covariantiematrix dan het EM-algoritme, indien de data redelijk gecorreleerd zijn en alle variabelen dezelfde varianties hebben. Echter, het REM-algoritme indiceert welke combinaties van variabelen met ontbrekende en geobserveerde data slechte schattingen van de (onbekende) ontbrekende data opleveren. Wanneer deze casus weggelaten worden, ontstaan nauwkeuriger schattingen van de covariantiematrix door het REM-algoritme.

6 Classificaties gebaseerd op met ruis vervuilde kenmerkwaarden

In hoofdstuk 6 wordt de invloed van meetruis op de classificatie van een casus geanalyseerd. In het hoofdstuk wordt de kwaliteitsmaat *robuustheid* voorgesteld die gebaseerd is op ideeën van Brender *el al.* De robuustheid van een classificatie is de waarschijnlijkheid dat de aan een casus toegekende klassenlabel niet zou veranderen indien men de classificatie had gebaseerd op de (onbekende) ruisvrije kenmerkwaarden. Een vooronderstelling hierbij is dat de meetruis Gaussisch is met een gemiddelde 0 en die ongecorreleerd is met de echte kenmerkwaarden. Een formule voor de robuustheid van een classificatie wordt gegeven.

In de praktijk blijkt dat het moeilijk is de robuustheid van een classificatie te schatten wanneer de dichtheidsfunctie van de ruisvrije kenmerkwaarden onbekend is. Daarom worden twee benaderingen voorgesteld waarbij men deze functie niet hoeft te kennen. Deze twee benaderingen veroorzaken een systematische fout in de schattingen van de robuustheid. De grootte van de geïntroduceerde fout wordt geanalyseerd in twee speciale situaties waarin een kenmerk unimodaal of bimodaal Gaussisch verdeeld is en de casus geclassificeerd zijn in een van twee klassen. Het blijkt dat een van de benaderingen een kleinere relatieve fout in de robuustheid geeft. De tweede benadering geeft grotere fouten, maar is rekentechnisch eenvoudiger.

Een simulatie-experiment laat zien hoe vaak een kenmerk opnieuw moet worden gemeten om een robuuste classificatie te garanderen. Het opnieuw meten van een

met ruis vervuild kenmerk is alléén zinvol wanneer weinig extra metingen vereist zijn om een robuuste classificatie te waarborgen. Indien de robuustheid van een casus zodanig laag is dat een kenmerk vaak opnieuw gemeten moet worden om een robuuste classificatie te verkrijgen, kan men beter afzien van het opnieuw meten van dit kenmerk. Het begrip *remeasuring interval* wordt geïntroduceerd. Deze intervallen geven aan wanneer het opnieuw meten van een kenmerk zin heeft.

7 Algemene conclusies

De methoden en technieken die worden voorgesteld in dit proefschrift zijn exploratief onderzocht in een aantal experimenten. In hoofdstuk 7 wordt bediscussieerd in welke mate deze methoden de ontwikkeling, verificatie en validatie van (toepassingsgerichte) neurale netwerken ondersteunen. De klinische toepasbaarheid van kennissystemen in het algemeen en neurale netwerken in het bijzonder wordt hierbij ook kort besproken. Het blijkt dat de introductie van deze systemen in de kliniek direct interfereert met de werkprocessen van de arts. Men verwacht dat deze systemen hun grootste bijdrage kunnen leveren bij het verwerken van grote hoeveelheden gegevens zonder dat daarbij hogere orde interpretaties nodig zijn. In vervolgonderzoek dient de waarde van de voorgestelde methoden en technieken voor ontwikkeling en evaluatie van neurale netwerken voor medische toepassingen vastgesteld te worden.

Zusammenfassung

1 Ziel dieser Arbeit

Diese Dissertation präsentiert Methoden und Techniken, die die Anwendung neuronaler Netze in der Medizin zu beschleunigen (helfen) können. Die Forschung über neuronale Netze begann als Zweig der Neurologie, da ein neuronales Netz typischerweise aus einer Menge miteinander verbundener Neuronen besteht. Die Neuronen, die als Verbindungs- oder Knotenpunkte zwischen Nervenfasern arbeiten, antworten auf eingehende Signale durch ein Ausgabesignal, eine Aktivierung. Diese Aktivierung steht in einem funktionalen Zusammenhang zu den Eingabesignalen, die zumeist durch Gewichte modifiziert werden. Seit der Einführung neuronaler Netze durch McCulloch und Pitts im Jahr 1943 erfuhr diese wissenschaftliche Feld eine Erweiterung hin zu den Kognitionswissenschaften, der künstlichen Intelligenz, der Statistik und Entscheidungstheorie, der Signaltheorie und den Ingenieurwissenschaften. Eine Vielzahl von Anwendungen neuronaler Netze in den Wirtschaftswissenschaften, der Informatik, der Telekommunikation und der Medizin belegt diese Entwicklung. Kapitel 1 enthält einen kurzen Überblick über Anwendungen neuronaler Netze, die für die Entscheidungsunterstützung in der Medizin entwickelt wurden.

Klinischer Anwendungen neuronaler Netze sind nicht unproblematisch aufgrund ihrer "black-box"-Eigenschaft. Es ist – zur Zeit – verhältnismäßig schwierig, das in den Gewichten des Netzes codierte Wissen eines trainierten neuronalen Netzes greifbar zu machen, da ein solches Netz eine nichtlineare Abbildung zwischen dem als Eingabe bereitgestellten Merkmalsraum und den als Ausgabe erwarteten Klassen realisiert. In dieser Arbeit werden deshalb Techniken vorgestellt, die die Eigenschaften eines trainierten neuronalen Netzes charakterisieren. Damit kann die Entwicklung neuronaler Netze beschleunigt und ihr Verhalten verifiziert beziehungsweise validiert werden. Die Diskussion des "missing value" Problems schließt die Arbeit ab.

2 Ein neuronales Netz realisiert eine Abbildung

Die allgemein verwandte Beschreibung für einen Klassifikator ist die mathematische Abbildung eines n-dimensionalen Eingaberaumes auf einen c-dimensionalen Ausgaberaum, N: $\mathbb{R}^n \rightarrow \mathbb{R}^c$, wobei n die Anzahl der Merkmale und c die Anzahl der zu unterscheidenden Klassen bezeichnet. Neuronale Netze können prinzipiell Kombinationen qualitativer und quantitativer Merkmale verarbeiten. Die c Ausgabeklassen können Entscheidungen, z.B. Therapien oder Diagnosen repräsentieren. Die Abbildung wird durch das neuronale Netz, genauer gesagt durch die gewichteten Verbindungen zwischen der Eingabe-, der verdeckten und der Ausgabeschicht realisiert. Dabei findet die Einstellung der Gewichte durch ein Training mit Hilfe eines Lerndatensatzes statt.

3 Beurteilung der Ausgabe eines Klassifikators

In Kapitel 2 werden Metriken definiert, mit denen die Leistungsfähigkeit eines trainierten neuronalen Netzes beschrieben werden kann. Die Klassifikationsleistung wird gemessen, indem das Netz eine Menge von Testfällen – den Testdatensatz – klassifizieren muß. Hierbei ist für jeden Fall die korrekte Klassenzugehörigkeit bekannt. Die Ergebnisse werden sodann in einer Kontingenztafel dargestellt, daß sie die Grundlage für die Anwendung bestehender Maße zur Leistungscharakterisierung neuronaler Netze sind. Diese Maße werden diskutiert und erweitert. Obwohl sie für neuronale Netze entwickelt wurden, sind sie nicht auf diese beschränkt, sondern können zur Charakterisierung jedes Klassifikators herangezogen werden, dessen Ergebnisse durch einen Kontingenztafel beschrieben werden können. Die Maße umfassen die Korrektheit (correctness) – die Menge aller korrekt klassifizierten Fälle – und die Bedeckung (coverage) – die Menge aller Fälle, zu denen das neuronale Netz eine Klassenzugehörigkeit vergeben kann. Die fehlklassifizierten Fälle werden durch die Maße Verschiebung (bias) und Streuung (dispersion) charakterisiert. Für einige Maße ist der Standardfehler und das Konfidenzintervall beschrieben. An Hand von Labordaten aus einer Studie über Schilddrüsenerkrankungen wird die praktische Anwendung und Aussagekraft der Maße illustriert.

4 Messung der Bedeutung von Merkmalen

Die Kapitel 3 und 4 der Arbeit beschäftigen sich mit der Messung der Bedeutung einzelner Merkmale des Eingaberaumes für die Gesamtklassifikationsleistung des neuronalen Netzes. Hierbei werden zwei Ziele verfolgt. Einerseits soll ein Einblick gewonnen werden, welche Merkmale wichtig sind für die Vergabe des richtigen Klassenlabels bezogen auf eine oder mehrere Klassen. Andererseits soll eine Abstufung der Merkmale hinsichtlich ihres Gesamtbeitrages realisiert werden, um unwichtige Merkmale zu eliminieren, sei es um sie durch bessere zu ersetzten oder um die Netzstruktur zu reduzieren.

In Kapitel 3 werden verschieden Suchtechniken zur Merkmalsauswahl diskutiert. Die Rückwärts-Suche erweist sich dabei als geeignet. Basierend auf der mathematischen Analyse des Bayes-Klassifikators wird eine Metrik der Diskriminationsleistung eines Attributs eingeführt. Diese Metrik wird verwendet, um die Merkmale des Eingaberaumes an Hand eines Testdatensatzes anzuordnen. Eine Summierung dieser Werte für alle Merkmale über alle Fälle des Testdatensatzes wird verglichen mit Friedman's zweiseitiger Varianzanalyse. Merkmale mit hoher mittlerem Rang erweisen sich als unwichtig für das Netz, während solche mit niedrigem mittleren Rang sich als äußerst "einflußreich" auf die Klassifikationsleistung des Netzes erweisen. Die Nützlichkeit dieses Ansatzes wird anhand verschiedener künstlicher Klassifikationsprobleme erläutert. Dabei wird deutlich, daß der Ansatz die Merkmale korrekt anordnet, unabhängig davon ob die Merkmale klassenbezogene Abhängigkeiten aufweisen oder nicht. Eine Anwendung auf Röntgenbilder fokaler Knochenläsionen zur Differenzierung von vier Texturen wird als Abschluß des Kapitels vorgestellt.

Im vierten Kapitel wird ein mathematischer Rahmen für vier Maße zur Charakterisierung der Merkmalseigenschaften für einen Bayes-Klassifikators festgelegt. Jede Metrik erlaubt die Berechnung einer unteren Schranke für den zusätzlichen Beitrag eines Merkmals auf die Leistung eines statistischen Klassifikators. Die Maße charakterisieren den Einfluß (*influence*) und die Austauschbarkeit (*replaceability*) eines Merkmals. Der Einfluß ist die Wahrscheinlichkeit, mit der ein Merkmal die Zuordnung eines Klassenlabels verändern kann, wenn die Werte aller anderen Merkmale eingefroren werden. Die Austauschbarkeit beschreibt den erwarteten Leistungsabfall, wenn ein Merkmal durch seinen Mittelwert ersetzt wird. Jedes dieser Maße wird durch die Definition einer zugeordneten Metrik handhabbar gemacht. Die Berechnung von drei von vier Metriken erfordert die Identifikation merkmalsbezogener Entscheidungsgrenzen. Die Entscheidungsgrenzen eines gegeben Merkmals hängen von den Werten der $n-1$ Merkmale ab und müssen für jeden Fall gesondert ermittelt werden. Diese Identifikation wird durch eine polynomiale Approximation realisiert, die auf der Taylor-Reihenentwicklung der Ausgabe des neuronalen Netzes als Funktion eines gegeben Merkmals basiert.

Anschließend wird eine Reduktionsprozedur, die als LMS-pruning bezeichnet wird, vorgestellt. Bezogen auf ein Merkmal erfolgt die Reduktion durch Entfernen aller Links zwischen dem für dieses Merkmal zuständigen Eingabeneuron und der verdeckten Schicht und dem Modifizieren aller verblieben Links zwischen Eingabeschicht und Hidden Layer so, daß die Klassifikationsleistung mit $n-1$ Merkmalen identisch ist zu der Variante mit n Merkmalen.

In Versuchen mit künstlichen Klassifikationsaufgaben werden die vier Maße hinsichtlich ihre Leistungsfähigkeit zur Anordnung der Merkmale nach Wichtigkeit untersucht und verglichen. Diese Versuche zeigen, daß die Austauschbarkeit (replaceability) das beste Anordnungsmaß darstellt. Die Experimente zeigten weiterhin, daß ein neuronales Netz mit einer Leistung dicht an der eines Bayes-Klassifikators durch Anwendung des LMS-prunings mit seiner Leistung dicht an der maximalen Korrektheit bleibt.

5 Abschätzung fehlender Daten

In Kapitel 5 wird eine Methode zur iterativen Abschätzung fehlender Daten vorgestellt. Statistische Klassifikatoren so wie neuronale Netze benötigen einen vollständigen Eingabevektor, um ein Klassenlabel vergeben zu können. Dies verhindert die Anwendung solcher Klassifikatoren in Gebieten, wo unvollständige Daten häufig auftreten. Verschieden Ansätze zur Abschätzung solcher fehlenden Daten wie der EM-Algorithmus oder die "multiple imputation" werden diskutiert. Um mit einigen Unzulänglichkeiten dieser Verfahren umgehen zu können wird ein wiederkehrendes auto-assoziatives Verfahren zur Abschätzung der fehlenden Daten vorgeschlagen. Zunächst wird die Situation analysiert, in der das auto-assoziative Verfahren mit vollständigen Trainingsdaten trainiert wird. Danach wird eine Abschätzung der fehlenden Daten mit wiederkehrendem Modus des auto-assoziativen Verfahrens versucht (recurrent auto associator). Danach werden Konvergenzbedingungen untersucht und es wird gezeigt, daß eine Konvergenz nur dann möglich ist, wenn die

Anzahl der verdeckten Neuronen des Auto-Assoziators kleiner oder gleich der Anzahl der beobachteten Werte eines unvollständigen Falls ist.

Der "recurrent auto associator" ist eingebettet in den "recurrent expectation maximization" (REM) Algorithmus, ein iterativer Ansatz zur Abschätzung fehlender Daten in einer Menge von Fällen. In einem Satz von Experimenten wird die residuale Varianz der Vorhersage durch den Auto-Assoziator verglichen mit der residualen Varianz einer multiplen Regression. Sowohl der REM- als auch der EM-Algorithmus werden im Hinblick auf ihre Fähigkeit zur Abschätzung der fehlenden Daten und der Kovarianzmatrix aus dem unvollständigen Datensatz untersucht und verglichen. Die Experimente zeigen, daß der REM-Algorithmus eine schlechtere Abschätzung der fehlenden Daten liefert als die multivariate Regression. Ferne schätzt der REM-Algorithmus die Kovarianzmatrix geringfügig schlechter ab als der EM-Algorithmus, wenn die gegeben Daten korreliert sind und alle Variablen gleiche Varianz haben. Durch den REM-Algorithmus kann aber die Aussage getroffen werden, in welchen Kombinationen von fehlenden und vorhandenen Merkmalen eine Abschätzung der fehlenden Daten schlecht möglich ist. Diese Kombinationen können dann gezielt unterdrückt werden, was zu einer Verbesserung der Abschätzung der Kovarianzmatrix führt.

6 Klassifikation mit verrauschten Merkmalen

In Kapitel sechs wird der Einfluß von meßtechnisch bedingtem Rauschen auf die Klassifikation eines Falls untersucht. Ausgehend von Ideen von *Brender et. al* wird das Qualitätsmaß Robustheit (robustness) spezifiziert. Die Robustheit einer Klassifikation ist die Wahrscheinlichkeit, daß das einem Fall zugewiesene Klassenlabel sich nicht verändert, wenn die Klassifikation auf den wahren (aber unbekannten) Werten durchgeführt wird. Dabei wird vorausgesetzt, daß das Meßwertrauschen als mit den Merkmalen unkorreliertes Gaußrauschen mit Mittelwert 0 angenommen werden kann. Auf dieser Basis wird eine Formel zur Berechnung der Robustheit einer Klassifikation angegeben.

In der Praxis ist es schwierig, die Robustheit einer Klassifikation zu schätzen, wenn die Verteilungsdichtefunktion der unverrauschten Merkmale unbekannt ist. Es werden deshalb zwei Näherungen vorgeschlagen. Der durch diese beiden Näherungen bewirkte systematische Fehler wird für die Spezialfälle untersucht, daß ein Merkmal einer unimodalen oder bimodalen Verteilung folgt und in ein oder zwei Klassen eingeordnet wird.

Eine Simulation illustriert, wie oft ein Merkmal gemessen werden muß, bis eine robuste Klassifikation vorliegt. Dabei wird deutlich, daß eine Wiederholung der Messung eines Merkmals nur Sinn macht, wenn wenige erneute Messungen zu einer hinreichend robusten Klassifikation führen. Umgekehrt wird deutlich, daß bei zu niedriger Robustheit die Anzahl der notwendigen Messungen sehr hoch wird, um eine hinreichend verläßliche Klassifikation zu erzielen. Es wird deshalb der Begriff der Messwiederholungsintervalle eingeführt. Diese Intervalle kennzeichnen, ob die erneute Messung eines Merkmals Sinn macht oder nicht.

7 Abschlußbemerkung

Die in dieser Dissertation entwickelten Methoden und Techniken werden an Hand verschiedener Experimente untersucht. In Kapitel 7 wird diskutiert, in welchem Umfang sie bei Entwicklung, Verifikation und Validierung neuronaler Netze helfen können. Die Möglichkeit, generell wissensbasierte Systeme und speziell neuronale Netze in der Medizin anzuwenden, wird ebenfalls angesprochen. Dabei wird deutlich, daß die Einführung solcher Systeme direkt mit dem Arbeitsbereich klinisch tätiger Ärzte in Konkurrenz treten kann. Das größte Potential ist deshalb in der "low-level" Informationsverarbeitung zu sehen. Es ist Gegenstand weiterer Forschung, den Wert der hier vorgestellten Methoden für die Entwicklung und Evaluation neuronaler Netze in medizinischen Anwendungen zu untersuchen.

Resume

1 Afhandlingens emneområde

I denne afhandling præsenteres en række metoder og teknikker der vil kunne fremme klinisk anvendelse af neurale netværk. Kunstige neurale netværk blev oprindeligt introduceret indenfor neurologi. Et kunstigt neuralt netværk består af et antal med hinanden forbundne neuroner. Hver neuron udgør et knudepunkt for et antal nervebaner. Gennem hver nervebane sendes et inputsignal til neuronen, hvilket resulterer i et outputsignal, der er funktionelt afhængigt af de modtagne input. Hvert input modificeres af en specifik vægt, et reelt tal. Siden neurale netværk blev introduceret i 1943, har de spredt sig til andre discipliner såsom kognitiv videnskab, kunstig intelligens, statistisk regressions- og beslutningsteori, signalbehandling samt et antal ingeniørdiscipliner. Neurale netværk er blevet udviklet til at løse bestemte opgaver indenfor økonomi, datalogi, telekommunikation og medicin, for bare at nævne nogle områder. I kapitel 1 gives et kort overblik over medicinske anvendelse af neurale netværk.

Det er problematisk at anvende neurale netværk til at løse opgaver indenfor det medicinske område på grund af deres black-box problem, som består i, at kvaliteten af den gennem træning af netværket indlærte viden meget vanskeligt lader sig bedømme. Netværkets forbindelser etablerer en ikkelineær mapning fra et featurerum (input) til et klasserum (output), der parametriseres af netværkets vægte. I afhandlingen udvikles en række teknikker, med det formål at beskrive egenskaber ved et trænet neuralt netværk. Derved understøtter disse teknikker iterativ udvikling, verifikation og validering af neurale netværk. Med det formål yderligere at fremme anvendelsen af neurale netværk, bliver der i kapitel 4 udviklet en metode til estimation af manglende data.

2 Et neuralt netværk udfører en afbildning

Den mest generelle specifikation af en klassifikator er en matematisk afbildning fra et n-dimensionals inputrum til et c-dimensionals outputrum, N: $\mathbb{R}^n \rightarrow \mathbb{R}^c$, hvor n udgør antallet af features eller attributter og c antallet af klasser, man ønsker at skelne imellem. En fordel ved neurale netværk er, at deres input kan være både af kvantitativ og af kvalitativ art. De c klasser består typisk af beslutninger som for eksempel diagnoser eller behandlinger. Afbildningen fra features til klasser foretages at det trænede neurale netværk, mere specifikt af de vægtede forbindelser mellem inputlaget, det skjulte lag og outputlaget. Vægtene bliver trænet gennem en læreproces, hvorunder netværket lærer at klassificere flest mulige træningsvektorer korrekt.

3 Kvalitetsvurdering af en klassifikators output

I kapitel 2 defineres en række metriker, som karakteriserer forskellige egenskaber ved et neuralt netværk i relation til dets præsterede output. Dette sker ved at lade et neuralt netværk klassificere en række testvektorer, for hvilke deres sande klassetilhørselsforhold er kendt. Ved at rubricere klassifikationsresultatet i form af kontingenstabel, kan forskellige egenskaber ved et netværk karakteriseres. Først diskuteres egnetheden af et antal eksisterende kvalitetsmetriker til at karakterisere et neuralt netværk. Alle disse metriker baserer sig på information, der er tilstede i en kontingenstabel. Derefter introduceres en række nye metriker. Selvom disse metriker alle er defineret med det formål for øje at beskrive egenskaber ved et trænet neuralt netværk, er metrikerne defineret så generelt, at de ligeledes kan anvendes til at beskrive egenskaber ved andre typer af klassifikatorer, blot disses præstationer kan karakteriseres ved hjælp af en kontingenstabel. Metrikerne inkluderer korrekthed – andelen af korrekt klassificerede vektorer – og dækningsgrad – andelen af vektorer der kan klassificeres af et neuralt netværk. De misklassificerede vektorer karakteriseres af metrikerne *bias* og *spredning*. Såkaldte standard errors og konfidensintervaller er angivet for nogle af metrikerne. Metrikernes anvendelighed undersøges gennem en række eksperimenter, hvor neurale netværk trænes til at klassificere Thyreodeapatienter.

4 Bedømmelse af attributbidrag til en klassifikator

Kapitel 3 og 4 behandler begge bedømmelse af det marginale bidrag, en attribut yder til et neuralt netværks præstation (korrekthed). Sådan en bedømmelse tjener dels det formål at opnå indsigt i hvilke attributter er vigtigere for at kunne klassificere en vektor korrekt, dels det formål at kunne ordne attributterne i henhold til det bidrag, de yder til en neuralt netværks præstation. På basis af denne viden kan man beslutte ikke at benytte attributter med et forsvindende lille bidrag til netværkets præstation.

I kapitel 3 diskuteres indledningsvis forskellige strategier til udvælgelse af attributter såsom forward, backward og Branch-and-Bound søgning. Backward søgning findes velegnet som udvælgelsesstrategi. På basis af en matematisk analyse af en minimal fejlrate (bayesiansk) klassifikator defineres en metrik, der karakteriserer det marginale bidrag, en attribut yder til klassifikatorens præstation. Denne metrik bruges til at rangordne de n attributter for hver vektor i et testsæt i henhold til deres bidrag. For hver attribut summeres over alle vektorer de til denne attribut tilkendte rang. Disse n summer, der er et udtryk for hver enkelt attributs bidrag, sammenlignes med Friedman's tovejs variansanalyse. Attributter med en gennemsnitlig høj rang har et relativt ringe bidrag til det neurale netværks præstation, hvorimod attributter med en gennemsnitlig lav rang bidrager mere til netværkets præstation. Igennem en række eksperimenter med kunstigt generede data undersøges det, hvorvidt den introducerede metode resulterer i en korrekt rangordning af attributterne. Disse eksperimenter indicerer, at metoden resulterer i den korrekte rangordning af attributterne, når den anvendes på neurale netværk, der er trænet til

at løse klassifikationsproblemer med uafhængige såvel som afhængige attributter. Metoden benyttes også i anvendelsesøjemed til at rangordne attributter, der er vigtige for at kunne skelne forskellige typer af textur i røntgenbilleder af blandt andet knogle tumorer.

I kapitel 4 udvikles et matematisk rammeværk, indenfor hvilket fire forskellige mål for det marginale bidrag af en feature bliver udledt for en minimal fejlrate-klassifikator. Hvert at disse featuremål gør det muligt at beregne en absolut undergrænse for det marginale bidrag, en feature yder til præstationen af en statistisk klassifikator. Disse mål karakteriserer den *indflydelse* en feature har samt dens *substituerbarhed*. Indflydelse er i denne kontekst defineret som sandsynligheden for, at en feature er i stand til at ændre klassifikationen af en vektor, mens de observerede værdier af de resterende features holdes konstante. Substituerbarhed af en feature defineres som den forventede reduktion af netværkets korrekthed, der optræder, når denne feature erstattes af sin betingede middelværdi.

Hvert af de fire featuremål operationaliseres ved en featuremetrik. For at kunne beregne tre af disse featuremetriker, kræves information om de grænser, der for en feature skiller de forskellige klasser. Disse betingede klassegrænser for en specifik feature afhænger af de $n-1$ featureværdier og må derfor identificeres for specifikt hver vektor. Til dette formål benyttes en polynomapproksimation, der er baseres på en Taylorrækkeudvikling af et neuralt netværks output som funktion af den givne feature. Den sidste metrik kan beregnes uden kendskab til de betingede klassegrænser.

En metode til beskæring af inputneuroner kaldet *LMS-pruning* bliver introduceret. LMS-pruning består i at fjerne forbindelserne mellem den inputneuron, man vil bortelimenere, og alle neuroner i det skjulte lag samt at ændre alle vægte mellem de resterende inputneuroner og de skjulte neuroner. Vægtene ændres således, at de beskårne netværk klassificerer træningssættet (baseret op $n-1$ features) identisk med et netværk, baseret på alle n features, men hvor værdien af den bortelimenerede inputneuron er erstattet af sin betingede middelværdi.

I en række eksperimenter baserede på kunstige klassifikationsproblemer sammenlignes de fire featuremetriker med hensyn til deres rangordning af de n features. Disse eksperimenter viser, at featuremetriken for substituerbarhed resulterer i den bedste ordning, nemlig den der er tættest på den sande (Bayesianske) rangordning af de n features. Eksperimenterne viste endvidere, at for neurale netværk, hvis korrekthed er næsten optimal i Bayesiansk forstand, resulterede LMS-pruning i beskårne netværk, hvis korrektheder forblev næsten optimale.

5 Estimation af manglende data

I kapitel 5 introduceres en metode til iterativ estimation af manglende data. Statistiske klassifikatorer såsom neurale netværk kan ikke anvendes på inkomplette vektorer, da netværkets output så ikke kan beregnes. Dette hæmmer blandt andet klinisk anvendelse af sådanne klassifikatorer, da man her ofte konfronteres med inkomplette data. Der eksisterer et antal metoder til behandling af datasæt med manglende værdier. To sådanne metoder er EM-algoritmen og Multipel Inputation;

fordele og ulemper ved begge diskuteres i dette kapitel. For at modvirke nogle af deres ulemper, foreslås det at benytte et cyklisk autoassociativt netværk til at estimere de manglende værdier. Først bliver den situation analyseret, hvori dette netværk trænes med komplette vektorer. Det bevises, at konvergens kun kan optræde, når antallet af neuroner i de skjulte lag er mindre eller lig antallet af observerede værdier i en inkomplet vektor.

Det cyklisk autoassociative netværk udgør byggestenen i REM-algoritmen, der er en iterativ metode til estimation af manglende værdier i et datasæt. I et antal eksperimenter sammenlignes den residualvarians, hvormed de manglende værdier forudsiges ved hjælp af det cyklisk autoassociative netværk, med den residualvarians, der opnås når de samme værdier forudsiges ved hjælp af multivariat lineær regression. REM- og EM-algoritmerne sammenlignes ligeledes, både hvad angår residualvarians af de estimerede manglende værdier og hvad angår deres estimationer af kovariansmatricen. Disse eksperimenter viser, at det cyklisk autoassociative netværk resulterer i ringere estimater af de manglende værdier end multivariat regression. REM-algoritmen estimerer kovariansmatricen lidt ringere end EM-algoritmen, når de inkomplette variable udviser en pæn korrelation med de komplette variable. Imidlertid kan REM-algoritmen indicere, hvilke kombinationer af variable med manglende og observerede værdier, der resulterer i estimater af de manglende værdier, som er behæftede med stor residualvarians. Udeladelse af sådanne vektorer, medfører et forbedret estimat af kovariansmatricen.

6 Indflydelse af støj på klassifikationsresultatet

I kapitel 6 analyseres den øgede usikkerhed i henhold til en klassifikation, som følger af, at attributter måles med støj. Baseret på en idé af Brender *et al.* defineres et kvalitetsmål kaldet *robusthed*. Robustheden af en klassifikation af en vektor er sandsynligheden for, at vektoren ville blive klassificeret anderledes, hvis klassifikationen baserede sig på de sande (støjfrie) men ukendte attributværdier. Det antages, at støjen er normalfordelt med en middelværdi på 0 og at støjen er ukorreleret med de sande attributværdier. En formel for robusthed specificeres for en statistisk klassifikator.

I praksis viser robusthed sig svær at estimere, når fordelingsfunktionen for de støjfrie attributter er ukendt. Derfor foreslås to approksimationer til robustheds-målet, der ikke baserer sig på denne fordelingsfunktionen. Disse approksimationer forårsager dog en bias af robusthedsestimatet, som analyseres for de to specialtilfælde, hvor en attribut er unimodal eller bimodal fordelt og man ønsker at skelne mellem to klasser.

Et simulationseksperiment illustrerer, hvor mange gange man er nødt til at måle en attribut (og udregne deres gennemsnit, hvilket reducerer indflydelsen af målestøj) for at opnå en tilstrækkelig robust klassifikation. Selvfølgelig er det kun meningsfuldt at genmåle en attribut, når få målinger er tilstrækkelige til at sikre en robust klassifikation. Hvis robustheden af en klassifikation er for lille, visers simulationerne, at det nødvendige antal genmålinger bliver så uforholdsmæssigt stort, at de med genmålingerne forbundne omkostninger bliver for store. Baseret på robustheds-

konceptet defineres såkaldte *genmålingsintervaller* for en attribut, der indicerer, hvornår genmåling er meningsfuld.

7 Generel konklusion

De i afhandlingen udviklede metoder og teknikker undersøges eksplorativt i eksperimenterne, som refereret i de enkelte kapitler. I kapitel 7 diskuteres det, hvorvidt de kan understøtte udvikling, verifikation og validering af neurale netværk. Klinisk anvendelse af vidensbaserede systemer og mere specifikt af neurale netværk bliver diskuteret. Der argumenteres for, at introduktion af sådanne systemer i klinikken griber direkte ind i lægernes arbejdsprocesser. Baseret herpå samt på det faktum, at man ikke har tilstrækkelig indsigt i deres egenskaber, konkluderes det, at neurale netværk har deres største potentiale indenfor lavniveau databehandling. Det er et emne for yderligere forskning at undersøge, hvorvidt de i denne afhandling præsenterede metoder og teknikker understøtter udvikling og kvalitetsvurdering af neurale netværk for medicinsk anvendelse.